International Journal for Housing Science and Its Applications
Publish August 3, 2025. Volume 46, Issue 3 Pages 868-884

https://doi.org/10.70517/ijhsa46365

The Application of Multimodal Information Fusion Algorithm in Teaching English Vocabulary in Colleges and Universities under the Perspective of Artificial Intelligence and the Network Security Guarantee Mechanism

Wei Dong[1,*]
[1] College of Foreign Language, Daqing Normal University, Daqing, Heilongjiang, 163712, China
Corresponding authors: (e-mail: Vivandongweiyy@163.com).

Abstract In the era of "Artificial Intelligence", the introduction of multimodal information fusion into vocabulary teaching is an important breakthrough in the reform of university English teaching in colleges and universities. In this paper, multimodal data are extracted from text, pictures and other domains, and the information of different modal data is fused through heterogeneous data fusion. Add positional coding and word vector embedding coding fusion operations in the information initialization stage, extract image features and text features, and send the information to the lexical model for fusion coding, use Transformer learning to decode the source utterance into the target utterance through the decoder, and use the Glove word vector model to realize the knowledge point vectorization operation in the knowledge point embedding layer. Design empirical analysis experiments to study the application effect of multimodal information fusion in English vocabulary teaching. The significance levels of the two classes of subjects in contextual discrimination and word selection by looking at pictures are 0.028 and 0.035 respectively, with the significance level less than 0.05, which indicates that the vocabulary learning method using multimodal information fusion algorithm is more effective in memorizing the words than the traditional mode. The network security mechanism is established, and the multimodal heterogeneous data operation security is evaluated through simulation experiments. The method in this paper can guarantee the data processing volume of 2~2.1Mb/s, and has high storage efficiency.

Index Terms Heterogeneous data fusion, Transformer learning, Glove word vector model, English vocabulary teaching

I. Introduction

Based on the development of digital network technology, the massive, personalized learning resources and teaching resources built into the digital teaching resources platform and the convenient and timely intelligent feedback system provide a convenient and intuitive measurement and analysis tool for teaching design and research [1], [2]. Based on this objective change, teachers need to take advantage of the situation and turn digital technology into an "enabler" of English teaching, promote the innovation of the traditional education model, and promote the transformation of the teaching space and the change of teaching methods [3]. The integration of digital technology into vocabulary teaching can greatly enrich the connotation and path of the English vocabulary classroom [4]. Numerous studies have shown that students' perception and interest are higher in comfortable, resource-rich, contextually diverse, and united learning environments as well as multi-modal mixed teaching [5]-[7]. Therefore, teachers should use the model of "PC platform resources" to integrate text, pictures, audio, video, animation, etc., drive vocabulary input and output with the help of multi-modal information means, strengthen students' perception and understanding of vocabulary from the perspective of multi-module and multi-modality, and build a student-centered connected, shared, autonomous, open, and appropriate vocabulary classroom [8]-[10].

In the 21st century, a large number of scholars have begun to introduce multimodal theories into the field of English language teaching and learning and have achieved remarkable results.Sutrisno et al [11] explored the effectiveness of multimodal literacy strategies integrated into English language learning environments, and found that the strategy could improve students' engagement and learning outcomes through multiple modes of communication (e.g., visual, audio, and digital technologies).Suwastini et al [12] reviewed the implementation and effectiveness of multimodal theory in English language teaching and found that the characteristics, benefits, and procedures of multimodal instruction, which varied according to teacher differences, had a positive impact on students' motivation, skills, and intelligence.Kummin et al [13] explored the effects of multimodal texts (including video, audio, and digital media), on students' English language skills and critical/creative thinking effects and found that ESL/L2 learning outcomes and student satisfaction can be improved through a multimodal approach adapted to different learning styles.Hafner et al [14] applied the Digital Multimodal Writing (DMC) model to English language teaching, which, unlike traditional writing, engages in multimodal communication while maintaining students' focus on language skills.Mu, H [15] stated that multimodal Teaching is based on the theoretical framework of multimodal discourse analysis, which applies images, sounds and gestures in actual classroom English teaching, so as to improve students' English proficiency and intercultural communication skills. Multimodal teaching methodology integrates a variety of cognitive strategies, according to the multimodal teaching strategies and cognitive modes, English vocabulary can be divided into different modules, and digital technology can be used to explain different modules in depth and vividly, and at the same time combined with the multimodal evaluation model to improve the effect of English vocabulary teaching [16].

In this paper, we mainly extract different modal data from two domains, text and visual features, and integrate these two modal data by means of heterogeneous data fusion, so that the model can better understand and utilize the data, and improve the utilization rate of the data. The extracted target data are sequentially subjected to encoding and decoding operations, and the Glove word vector model is used to realize the knowledge vectorization of the knowledge point embedding layer to complete the multimodal fusion of English vocabulary. A multimodal vocabulary teaching model is designed and the model is applied to English vocabulary teaching in colleges and universities, and a 16-week stage test is conducted for students. Taking NSSA technology as the basic framework, Endsley model and Agent theory are introduced respectively to propose a data vulnerability mining mechanism for multimodal data network, which guarantees the safe operation of the multimodal information fusion process.

II. Multimodal Information Fusion Algorithm Applied to English Vocabulary Teaching in Colleges and Universities

II. A.Relevant machine learning methods

II. A. 1) Multimodal information fusion methods

Multimodal information fusion is a multidisciplinary cross-cutting field, which mainly extracts data of different modalities from many different domains, such as text, pictures, sound, video, etc., and integrates the different modal data by means of heterogeneous data fusion, which enables the model to better understand and utilize the data, and improves the utilization rate of the data [17]. Multimodal technology fuses data information from different modalities, which makes the data information complementary to each other, increases the coverage of the data contained in the input data, improves the accuracy of the model on the prediction task, and enhances the robustness of the model. The main purpose of multimodal information fusion is to reduce the heterogeneous differences between different modal data while ensuring the specific semantic integrity of heterogeneous data. Information fusion can be mainly categorized into pre-fusion and post-fusion based on the feature extraction as well as the temporal order of modal fusion.

II. A. 2) Multimodal attention mechanisms

In the model, hidden states are computed as a self-attention and feed-forward system, while positional coding is used to show the regional relations of each word in the sentence, with the inclusion of a multi-head attention mechanism for parallelized processing. Due to the possibility of parallel processing, the processing speed of MT is accelerated, leading to better results than the lexical models based on recurrent neural networks. In MMT, the extracted image feature information is only used as auxiliary information to guide the lexical translation system for lexical translation. Therefore, visual feature information is not equally important compared to text. If the information is directly fused with positional coding and word vector embedding coding at the initial position, unnecessary noise will be generated.The image feature information is added to the encoder part of the overall architecture as shown in Fig. 1.Transformer consists of several sub-modules, each of which contains multiple heads of attention, feed-forward neural networks, and residual structures in the sub-modules. Each submodule processes each token and its previous input relations. At the beginning, the self-attention mechanism is used to construct the multi-head attention module. The method is to apply self-attention to the same input multiple times with separate normalization parameters. After the attention parameters are set once, they can be reused many times to solve parallel processing problems.

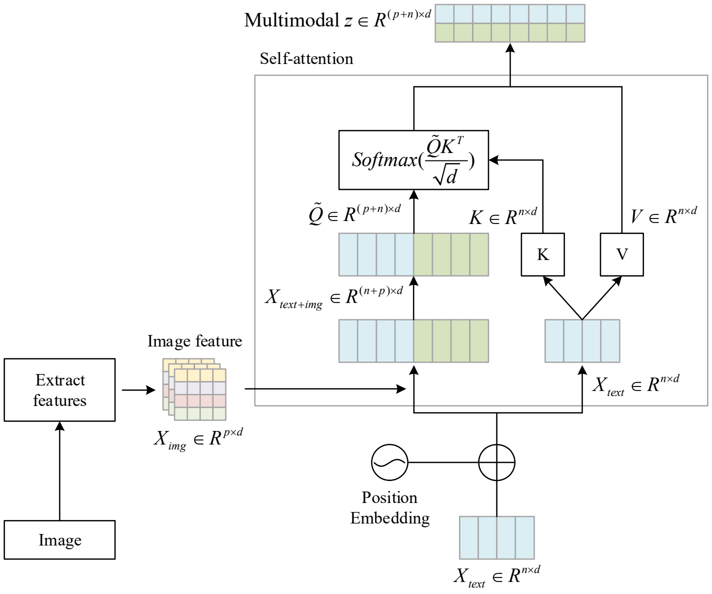

Multimodal $z \in R^{(p+n)\times d}$
Self-attention
$Softmax(\frac{\tilde{Q}K^T}{\sqrt{d}})$
$\tilde{Q} \in R^{(p+n)\times d}$
$K \in R^{n\times d}$
$V \in R^{n\times d}$
K
V
$X_{text+img} \in R^{(n+p)\times d}$
Image feature
Extract features
$X_{img} \in R^{p\times d}$
$X_{text} \in R^{n\times d}$
Position Embedding
Image
$X_{text} \in R^{n\times d}$

Figure 1: Adding visual information to self-attention coding

In addition, the model easily learns the correlation information between different heads by computing the multi-head attention method. During data processing, the self-attention module calculates and updates the weights based on the importance of the markers in the sequence. As mentioned before, the input screen columns are mapped to three matrices (K,Q,V). The dimensions of all three correlations are $\mathbb{R}^{d}\times h$, with d denoting the embedding dimension and h denoting the number of attention heads. The final output of each head will be fused by a linear connection. So taking text and image as the two parts of the information input as $x^{text} \in \mathbb{R}^{n\times d}$ and $x^{img} \in \mathbb{R}^{n\times d}$, we can get the result of multimodal attention as:

$$z_i = \sum_{j=1}^{n} \tilde{\gamma}_{ij}(x_j^{text} W^V) \tag{1}$$

where γ_{ij} is the weight matrix obtained from the softmax function. So it can be obtained:

$$\tilde{\gamma}_{ij} = softmax\left(\frac{(x_i^{(text+img)} W^Q)(x_j^{text} W^K)^T}{\sqrt{d}}\right) \tag{2}$$

where $z \in \mathbb{R}^{(n+p)\times d}$ is the hidden representation of the fused text and image. The decoder side receives the z generated at the encoder side and then generates the target sequence. The features of the extracted spatial image are not directly encoded in the model; they require the hidden representation of each word in the image to be computed by adjusting the attention. In each encoder layer, residuals as well as normalization layers are used and finally the decoder of the Transformer standard is connected.

II. A. 3) Visual feature extraction

In the lexical model, the extracted image features and text information are simultaneously fed into the lexical model for fusion coding. For image features, three models, VGG11, VGG19 and ResNet50, are selected for feature information extraction, and the features extracted by different models are analyzed to see whether they contain rich visual information and their contribution to the lexical translation task. The pre-trained VGG and ResNet15 models were fine-tuned to obtain the required image feature information.

II. B. Construction of English Vocabulary Teaching Model Based on Multimodal Information Fusion

II. B. 1) Source and target vision coders

As mentioned earlier, NMT uses self-attention to learn the relevance of each word to other tokens [18]. For MNMT, self-attention is then used to learn text-visual relations and save text-visual interaction information. In this paper, the encoder's multiple attention is used to capture source visual interaction information (SIV). Formally, $s^{text} = (s_1^{text}, \cdots, s_N^{text})$ is used to represent the textual information, and s_N^{text} is used to represent the N words of the textual sentence, and the outputs of the self-attention that incorporate the source visual interaction information are computed as follows:

$$z_i = \sum_{j=1}^{N} \alpha_{i,j} (s_j^{text} W_{enc}^{V}) \tag{3}$$

where $\alpha_{i,j}$ is the attention weight matrix computed by the softmax function:

$$\alpha_{i,j} = softmax\left(\frac{((s_i^{text} \oplus x^{img}) W_{enc}^{Q})(s_j^{text} W_{enc}^{K})^T}{\sqrt{d_k}} \right) \tag{4}$$

where x^{img} denotes the image information and $s_i^{text} \oplus x^{img}$ denotes the result of the superposition of s_j^{text} and x^{img}. Self-attention of fused visual information can dynamically provide potential adaptation of text to images to model source-visual interaction information. At the same time, target-visual interaction information (TIV) needs to be considered. The same operation as the encoder is used to obtain the text-visual interaction information at the target. The computation is as follows:

$$c_i = \sum_{j=1}^{M} \alpha_{i,j} (t_j^{text} W_{dec}^{V}) \tag{5}$$

$$\alpha_{i,j} = softmax\left(\frac{((s_i^{text} \oplus x^{img}) W_{dec}^{Q})(s_j^{text} W_{dec}^{K})^T}{\sqrt{d_k}} \right) \tag{6}$$

II. B. 2) English Vision Protocol Decoder

During the training process of NMT, the Transformer learns the probability of decoding a source utterance into a target utterance through a decoder as follows:

$$p_{s \to t}(t \mid s) = P(y \mid t^{text}, s^{text}) \tag{7}$$

For MNMT, visual information is fused into the encoder and decoder to simulate source and target visual interaction information. Thus, the potential representation of the encoder and decoder can simulate bilingual-visual interaction information. However, this situation still has a gap for bilingual-visual interaction information. The reason is that the same semantics of source and target words may align different contents of the image when fusing visual information to the encoder and decoder respectively. To solve this problem, a bilingual visual module is added to the decoder. This module is a combination of layered attention that fuses visual and textual features into the decoder and generates target sentences by adding visual information. Specifically, we first compute the context vectors $c_i^f (f = img)$ for each image, then project these context vectors of images and text into a common space and compute the alternative distribution of the projected context vectors:

$$e_i^f = \psi(s_i, t_i, c_i^f) \tag{8}$$

$$\beta_i^f = \frac{\exp(e_i^f)}{\sum_{r \in \{img, text\}} \exp(e_i^r)} \tag{9}$$

$$c_i = \sum_{r \in (img, text)} \beta_i^r W^r c_i^r \tag{10}$$

where ψ is a feedforward network, s_i is the i th source and target encoder hidden state, and t_i is a self-attention module that is used to merge image and text vectors. $W^{'}$ is a weight matrix which is used to compute the context vector. c_i is obtained from the image and text features. The final y is obtained as follows:

$$p_{s \to t}(t \mid s) = P(y \mid t_{test}, s_{text}, c) \tag{11}$$

where s denotes the standard neural network vocabulary translation prediction of different y, and the bilingual-visual consistent decoder requires the system to combine visually informative vocabulary to translate an increasing number of source sentences.

II. B. 3) Knowledge point embedding layer

In this paper, we use the Glove word vector model to realize the knowledge vectorization operation at the knowledge point embedding layer.Glove is a global logarithmic bilinear model that uses unsupervised learning to train word vectors. The model combines the advantages of both global matrix decomposition and local context windows, and unlike other models that train on the entire sparse matrix or a single context window in a large corpus, the model is trained directly on the word-word co-occurrence matrix, which efficiently utilizes the statistical information to generate a semantically-rich vector space.The most critical module in the Glove model is the The most critical module in the Glove model is the training of the co-occurrence matrix, and there are two ways to train the co-occurrence matrix in the model, namely, the symmetric window training without considering the order of words and the asymmetric window training considering the order of word context. In order to obtain high-quality word vectors, this paper adopts the asymmetric window approach to train the co-occurrence matrix of the model. The specific steps are as follows:

(1) The corresponding word list is generated by counting the number of occurrences of each word in the lexical corpus, and the words are sorted in the word list according to the number of occurrences of the word from high to low. c_i denotes the i th word that appeared, f_i denotes the number of times the i th word appeared, and n is the size of the word list, i.e., the number of different words in the lexical corpus.

(2) Set the sliding window size to w, traverse all the words in the corpus, and record the frequency of occurrence of the words in the fixed window to the left of the target word to generate the left co-occurrence matrix x^L, and denote the word in the left co-occurrence matrix in the i-row and j-column with x_{ij}^l.

(3) Use V^A to denote the low-dimensional word vector representation obtained from the training based on the left-hand side co-occurrence matrix, and train the model by the loss function J^A, which is computed as shown in Eq. (12):

$$J^A = \sum_{i,j=1}^{n} f(X_{ij}^L)((v_i^A)^T v_j^A + b_i^A + b_j^A - \log X_{ij}^L)^2 \tag{12}$$

As described in Eq. (12), where n is the size of the vocabulary list (co-occurrence matrix dimension $n*n$), v_i^A, v_j^A is the asymmetric low-dimensional word vector representation of words c_i and c_j, respectively, and b_i^A, b_j^A is the bias term corresponding to v_i^A, v_j^A, respectively, and $f(X_{ij}^L)$ is the weight function.

II. B. 4) Modal Fusion Layer

The main purpose of the modal fusion layer is to fuse features from the text vectors of the lexical text encoding layer, the picture vectors of the lexical picture encoding layer, and the knowledge point embedding vectors of the knowledge point embedding layer.

Firstly, the heterogeneous features are fused by a cross-modal cross-attention method based on the attention mechanism, which uses self-attention to establish a connection between the data of different modalities. The text feature vectors and image feature vectors obtained after the cross-modal cross-attention operation are input to the multi-head attention module for heterogeneous data feature fusion. The final output is a multimodal lexical representation vector fusing two modalities and three features that can represent the heterogeneous data features for the group scrolling task.

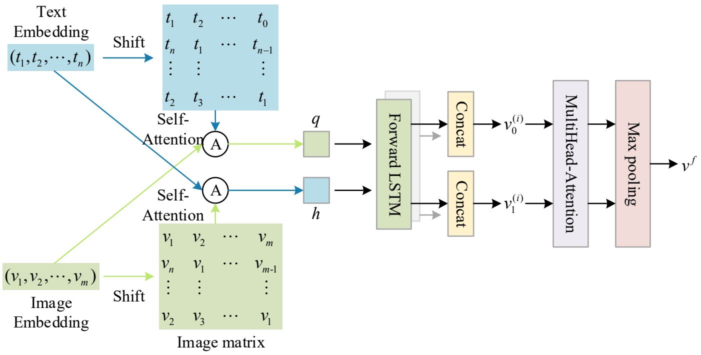

Text
Embedding
$(t_1, t_2, \cdots, t_n)$
Shift
t_1 t_2 $\cdots$ t_0
t_n t_1 $\cdots$ t_{n-1}
$\vdots$ $\vdots$ $\vdots$
t_2 t_3 $\cdots$ t_1
Self-Attention
A
Self-Attention
A
q
h
Forward LSTM
Concat
Concat
$v_0^{(i)}$
$v_1^{(i)}$
MultiHead-Attention
Max pooling
v^f
$(v_1, v_2, \cdots, v_m)$
Image
Embedding
Shift
v_1 v_2 $\cdots$ v_m
v_n v_1 $\cdots$ v_{m-1}
$\vdots$ $\vdots$ $\vdots$
v_2 v_3 $\cdots$ v_1
Image matrix

Figure 2: Schematic diagram of modal cross fusion

The text representation vector $sentence_{emb}$ that can represent the semantics of lexical text topics is obtained through the text encoding layer, the image feature vector $image_{emb}$ that can characterize the image information is obtained through the image encoding layer, and the knowledge point embedding vector kc_{emb} is obtained through the knowledge point embedding layer. The text representation vectors with the same text modality and the knowledge point embedding vectors are first summed to generate $text_{emb}$ as a uniform text vector representation of the input text data. The matrix representation generated by translating the text vector $text_{emb}$ and the image vector $image_{emb}$ is fed into the cross-modal attention module to exchange the heterogeneous information between the two modalities. Among them, the structure of cross-modal cross-attention module is shown in Fig. 2, where each element of the embedding vector obtained through the coding layer is firstly translated to the matrix representation of the corresponding modality, and then the attention operations are performed between the vectors and matrices of the different modalities respectively, so as to establish a connection between the data of the different modalities.

The specific way is to do the attention operation between the two modalities: text-based text-to-picture attention operation and picture-based picture-to-text attention operation, and finally get the text representation vector h fused with the picture features and the picture representation vector q fused with the text features, which are calculated as shown in Eq. (13) and Eq. (14):

$$h_i^k = CrossAttention(t_i^{k-1}, \{v_1^{k-1}, \ldots, v_m^{k-1}\}) \tag{13}$$

$$q_j^k = CrossAttention(v_j^{k-1}, \{t_1^{k-1}, \ldots, t_n^{k-1}\}) \tag{14}$$

Where t_i^k denotes the text vector, v_j^k denotes the picture vector, and CrossAttention is the attention operation.

The text feature vector h and picture feature vector q fused with the heterogeneous features are inputted into the bi-directional long short-term memory network Bi-LSTM to further establish the connection between the heterogeneous features. The forward hidden state quantity $\vec{h}_t$ at the moment of t in the bi-directional long short-term memory network is computed from the forward hidden state quantity $\overrightarrow{h_{t-1}}$ at the previous moment and the current input vector vector w_t, and the t The amount of moment reversed hidden state $\overleftarrow{h}_t$ is computed from the hidden state of the next moment $\overrightarrow{h_{t+1}}$ as well as the current input vector w_t. Thus the hidden state representation of each input vector can be computed by doing concatenate of the hidden states in both directions: $v_w = concatenate(\overrightarrow{h_w}, \overleftarrow{h_w})$. After the bi-directional long short-term memory network operation, we output the vector matrix v which contains the heterogeneous features. Finally, we obtain the position of each vector v_i in the vector matrix v and use one-hot position encoding to obtain the position encoded vector $pemb_i$, and then we concatenate the position of the vector v_i with the corresponding position encoded vector $pemb_i$. Do concatenate operation to obtain the feature vector v_c^i that incorporates the location information. Each feature vector v_c^i is used as an input to the multi-attention mechanism, and the feature vectors are aggregated using Maxpooling to generate a global multimodal representation vector that can finally characterize the heterogeneous vocabulary. The specific calculation of the multimodal representation vector is shown in Equation (15):

$$v^f = \max\{\mathrm{MultiHead}(\mathrm{LayerNorm}(v_c^i, v_c^i, v_c^i) + v_c^i)\} \tag{15}$$

where v_c^i is the feature vector incorporating location information, LayerNorm is the multilayer normalization technique, MultiHead is the multi-head attention mechanism, and v^f is the English vocabulary multimodal representation vector incorporating the knowledge point embedding vectors, vocabulary text vectors, and vocabulary appendage feature vectors of word selection from looking at a map in the input vocabulary, as the final global vector representation of the English vocabulary.

II. C. The Use of Multimodal Information Fusion Algorithm in Teaching English Vocabulary

II. C. 1) Constructing a multimodal teaching interaction model

Vocabulary is the source of linguistic meaning and the pillar on which the language system depends. As the smallest symbolic unit of language, vocabulary has conceptual, interpersonal and discourse functions, and the construction of its meaning is closely related to various modalities in life. Therefore vocabulary plays an extremely important role in understanding, expressing and transmitting ideas. Vocabulary and vocabulary knowledge can limit learners' comprehension of the language and their ability to read deeply.

In fact, vocabulary acquisition can be approached from the perspective of different cognitive strategies, integrating vocabulary constructions, sound representations, pictorial representations, image-visual associations and other cognitive strategies, which can be integrated into four modules: culture, vocabulary representation, vocabulary derivation and vocabulary expansion. On this basis, the multimodal information fusion model constructed is used to present the four modules in a reasonable way, and a multimodal evaluation system is established to help students acquire vocabulary and improve the efficiency of vocabulary learning. The model is a model of teaching interaction consisting of cognitive strategies and multimodality, i.e., the vocabulary cognitive strategy-multimodality teaching model.

II. C. 2) Lexical Derivation Module

The development of language vocabulary is a gradual process, which goes from nothing to something, and then from something to a lot, and its course of development tends to go through two stages: primary and derivational.

(1) Native Words

Primitive word is a key step in the development of English vocabulary from nothing to something. As the cornerstone of vocabulary evolution, its importance is self-evident. The emergence of native words is inextricably linked with their etymology. Different etymologies produce different root words in the process of language evolution, and the word formation mechanism of these root words is one of the important ways for the production of native words, therefore, it is crucial for English learners to have a deep understanding and knowledge of root words and affixes.

(2) Convergent Words

“Conjugation” is an important link in the process of English vocabulary from existence to multiplicity. This kind of word formation is an important method of language derivation, and is a very common phenomenon in the initial stage of language vocabulary expansion. Conjugation refers to the combination of two or more independent words or morphemes according to their respective meanings into a single word, and this kind of word formation is called conjugation.

II. D.The Effectiveness of Multimodal Information Fusion in Teaching English Vocabulary

II. D. 1) Research design

This paper focuses on the application effect of multimodal information fusion teaching mode in English vocabulary teaching in colleges and universities. This paper takes advantage of the opportunity of internship in a provincial key university in HLJ province to carry out an experiment to apply the multimodal information fusion teaching mode to daily vocabulary teaching to observe the application effect and collect experimental data. Fifty-five students from each of the two classes A and B of the first year, totaling 110 students, were selected as the experimental research subjects for vocabulary tests and questionnaires, and interviews were conducted with the teachers of the English teaching and research group of the first year and the individual students randomly selected to participate in the experiments of the two classes at the end of the experimental stage. The research was carried out by analyzing the pre-experimental vocabulary test data and post-experimental vocabulary test data of the research subjects, the questionnaire survey results and the records of the survey and interviews, and finally scientific conclusions were drawn.

II. D. 2) Research Objects

The research subjects of this experiment were 110 students in two parallel classes A and B of freshman year in a provincial key university in HLJ province. One of the classes, class A, was selected as the experimental class and the other parallel class, class B, was selected as the control class during the experimental study. A total of 110 students in the two classes will be subjected to the same pre- and post-experimental vocabulary tests and questionnaires, and the whole experiment will last for a period of 16 weeks of staged testing.

Table 1: Group statistic

/	Class	N	Mean	Standard deviation	Standard error of mean
Learning motivation	A	55	3.214	0.846	0.166
	B	55	2.469	0.816	0.114
Learning strategy	A	55	3.292	1.039	0.148
	B	55	2.312	0.834	0.169
The importance of lexical learning	A	55	3.498	0.863	0.196
	B	55	2.536	0.966	0.128
Traditional vocabulary teaching attitude	A	55	3.249	0.948	0.134
	B	55	2.816	0.632	0.063
Multi-modal information fusion teaching attitude	A	55	3.406	0.845	0.165
	B	55	2.432	0.746	0.096
Population	A	55	3.348	0.745	0.125
	B	55	2.505	0.436	0.063

II. D. 3) Analysis of experimental data

(1) Attitude toward learning English vocabulary

In order to compare whether there are significant differences between the two classes participating in the experiment in the various dimensions of the questionnaire, the author conducted an independent samples t-test on the questionnaire data collected from the two classes, and the results are shown in Tables 1 and 2.

Table 1 shows the descriptive statistical quantities of the two classes participating in the experiment in various dimensions, taking learning motivation as an example, it can be seen that class A in the learning motivation mean value is 3.214, class B in the learning motivation mean value is 2.469, from the size of the value can be seen that there is a slight difference between the two, but whether or not it is at a statistically significant level of difference, the results are shown in Table 2.

Table 2 shows the independent samples t-test, which shows that at a significance level of 0.05, the two classes have a t-value of 5.166 in terms of motivation to learn assuming equal variance, and a sig value of 0.000, which is less than 0.05, indicating that the variances are not chi-square. The results of independent samples t-test show that there is a significant difference between class A and class B in terms of motivation to learn and it can be seen from the size of the mean that class A is significantly higher than class B. Similarly it can be seen that there is a significant difference between the two classes in all other aspects and class A is significantly higher than class B. It can be concluded that the multimodal information fusion teaching model is conducive to increasing students' motivation, improving learning strategies, and positively affecting vocabulary learning.

(2) English vocabulary learning achievement

The scores were entered into SPSS (the highest and lowest scoring papers of the two tests in class A. By analyzing the differences between the two classes and within the classes in the two tests, the data can be obtained as shown in Table 3, where 1 is the vocabulary test score of the pre-test and 2 is the vocabulary test score of the post-test.

An independent samples t-test was conducted on the total English vocabulary test scores and the scores of each part of vocabulary, and the total scores of the pre-test and the results of the analysis of the dimensions showed that the scores of Class A were significantly higher than the scores of Class B. The level of significance in the total score (0.005), dictation (0.004), and English to Chinese translation (0.001) in the 0.01 level. The significance level of contextual discrimination as well as word choice by looking at the picture is 0.028 and 0.035 in 0.05 level respectively. The significance level of Chinese to English translation is less than 0.001. While in the post-test, on the total score and each part of the scores, class A scores significantly more than class B, and the significance level of both is less than 0.001. Before and after the two exams, the overall scores and each part of the scores of class A are slightly better than those of class B, and the advantage of class A in the post-test is even more obvious. This shows that the vocabulary learning method using the multimodal information fusion algorithm is more effective than the traditional mode in memorizing words, especially in the more difficult words. Meanwhile, from the data analysis, it can be seen that the difference of these five dimensions is bigger than that of the pre-test, especially in the aspect of choosing words by looking at the picture, the comparison data of class A and class B have the biggest difference compared with the other four dimensions.

Table 4 shows the pre- and post-test differences between high and low level students within the experimental class, grouping the group samples in terms of high and low achievement scores shows that in the pre-test there is a significant difference between the high and low groupings of the experimental class, with a difference in means ranging from 2.933-6.802, and in the post-test measurements there is a significant difference between the high and low groupings of the A class, with a difference in means ranging from 1.612-4. The difference between the two differences was smaller on the posttest than on the pretest. The dimension with the largest difference in for the pre-test was dictation and writing words, with a difference of 6.802. In the post-test, this difference narrowed significantly. Differences in the same test for students at different levels from when comparisons were made within the experimental class. In all experimental classes using the multimodal information fusion algorithm, the difference between the scores of students at higher and lower levels was larger in both exams, but in the posttest, the difference was narrowing. This suggests that the multimodal information fusion algorithm is more effective in memorizing more difficult words than simple words for poor students.

Table 2: Independent sample t test

/		Levene test of the variance equation				T test of the mean equation				
		F	Sig.	T	Df	Sig. (Double side)	MD	SD	95% confidence interval of the difference	
									Lower limit	Upper limit
Learning motivation	Assumed equal variance	0.148	0.796	5.166	108	0.000	0.866	0.164	0.515	1.278
	Suppose the variances are not equal			5.156	107.348	0.000	0.866	0.164	0.515	1.278
Learning strategy	Assumed equal variance	3.545	0.063	5.433	108	0.000	0.966	0.185	0.636	1.324
	Suppose the variances are not equal			5.366	100.653	0.000	0.966	0.183	0.636	1.266
The importance of lexical learning	Assumed equal variance	0.176	0.613	5.133	108	0.000	0.848	0.175	0.533	1.246
	Suppose the variances are not equal			5.123	107.325	0.000	0.848	0.175	0.538	1.248
Traditional vocabulary teaching attitude	Assumed equal variance	12.059	0.002	2.433	108	0.023	0.369	0.168	0.086	0.773
	Suppose the variances are not equal			2.421	93.348	0.015	0.369	0.167	0.085	0.778
Multi-modal information fusion teaching attitude	Assumed equal variance	1.536	0.269	6.096	108	0.000	0.912	0.153	0.631	1.269
	Suppose the variances are not equal			6.035	102.936	0.000	0.912	0.154	0.635	1.273
Population	Assumed equal variance	19.916	0.000	6.243	108	0.000	0.723	0.136	0.587	1.018
	Suppose the variances are not equal			6.315	86.612	0.000	0.723	0.136	0.588	1.036

Table 3: Class A and B are different in each dimension

/	Class	N	Mean	SD	T	P
Total score 2	A	55	84.395	5.966	15.269	0
	B	55	66.568	5.498		
Listen and write words 2	A	55	16.533	1.355	10.866	0
	B	55	13.648	1.539		
Look at the picture and choose the words 2	A	55	17.536	1.936	10.733	0
	B	55	13.355	1.736		
Chinese-English translation 2	A	55	16.321	1.688	9.948	0
	B	55	12.989	1.795		
English-Chinese translation 2	A	55	16.769	1.393	9.036	0
	B	55	13.268	1.863		
Context analysis 2	A	55	17.236	2.136	10.536	0
	B	55	13.308	1.348		
Total score 1	A	55	65.704	10.596	3.165	0.005
	B	55	56.067	19.266		
Listen and write words 1	A	55	12.936	2.384	2.835	0.004
	B	55	11.136	3.836		
Look at the picture and choose the words 1	A	55	12.798	3.218	2.109	0.035
	B	55	11.169	4.168		
Chinese-English translation 1	A	55	13.296	2.789	4.036	0
	B	55	10.489	4.063		
English-Chinese translation 1	A	55	13.348	2.348	3.069	0.001
	B	55	11.308	4.266		
Context analysis 1	A	55	13.326	2.063	2.158	0.028
	B	55	11.965	4.085		

In the course of the experiment, this paper conducted a stage test for students in weeks 1-16 of the experiment respectively. Unlike the pre-test and post-test, which examined students' comprehensive vocabulary use, the stage test mainly examined students' vocabulary mastery, i.e., whether students were able to write key vocabulary and high-frequency vocabulary according to the teacher's pronunciation, natural spelling, or normal memory. The vocabulary dictation was 20 words each time, and the scope of vocabulary dictation was the vocabulary that was being learned and had been learned in the experimental stage.

Through the recovery of the students' dictation books and counting the scores, the changes in the academic performance of the control class and the experimental class during the experimental period of the stage test are shown in Figure 3.

The average scores of the students of class A on the stage test in weeks 4, 8, 12 and 16 were 73.796, 77.841, 83.315 and 86.484 respectively. There was a gradual incremental increase in students' scores, with the most substantial improvement in the week 7 phase test scores compared to the previous six, as the corresponding phase test scores for students in class B were 75.779, 77.964, 79.314, and 80.403, respectively.The phase test scores for class B were similarly improving, but the degree of improvement was slow. In fact, after analyzing the specific data, it was found that the increase in Class B's scores came more from the progress of students whose scores were already in the upper middle range, while the scores of some students in the lower middle range were at a standstill or even declining instead of increasing. This suggests that unlike Group A, the traditional teaching approach, although it can also help students make some progress, cannot better take care of the progress of the poorer students, but instead leads to the phenomenon of achievement differentiation in the class.

Table 4: The difference between the high and low students in the laboratory class

/	Sort	N	Mean	SD	T	P
Total score 2	High level	27	90.786	1.248	12.349	0
	Low level	28	77.345	4.053		
Listen and write words 2	High level	27	17.248	1.436	3.615	0.002
	Low level	28	15.636	0.856		
Look at the picture and choose the words 2	High level	27	19.636	0.823	8.826	0
	Low level	28	15.736	1.436		
Chinese-English translation 2	High level	27	17.336	0.918	4.236	0
	Low level	28	15.289	1.648		
English-Chinese translation 2	High level	27	17.618	1.358	4.248	0
	Low level	28	15.736	1.038		
Context analysis 2	High level	27	18.948	1.439	6.936	0
	Low level	28	14.948	1.636		
Total score 1	High level	27	79.48	2.836	25.036	0
	Low level	28	54.735	2.598		
Listen and write words 1	High level	27	15.736	1.048	10.945	0
	Low level	28	11.026	1.266		
Look at the picture and choose the words 1	High level	27	16.625	1.537	13.085	0
	Low level	28	9.823	1.436		
Chinese-English translation 1	High level	27	16.123	1.536	10.169	0
	Low level	28	10.523	1.378		
English-Chinese translation 1	High level	27	15.648	1.236	9.759	0
	Low level	28	10.948	1.486		
Context analysis 1	High level	27	15.348	1.196	6.652	0
	Low level	28	12.415	1.345		

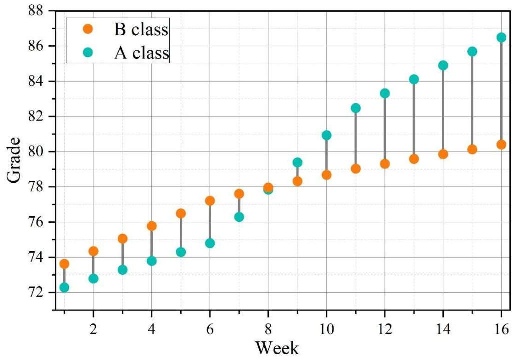


Figure 3: The study of the students' periodic tests was changed

III. Network security guarantee mechanism based on multimodal data fusion

III. A. Network Vulnerability Mining Based on Multimodal Data Fusion

III. A. 1) Fusing multimodal data features

The multimodal information fusion model of college English vocabulary designed in this paper operates by integrating different modal data through heterogeneous data fusion, and in this process, it is easy to generate the security problem of network data, therefore, in order to guarantee the security problem of the network operation scenario in this paper, it is necessary to fuse the extracted multimodal data features before carrying out the mining task.

III. A. 2) Mining network vulnerabilities

Generally, when conducting network security vulnerability mining, it is necessary to do it without affecting the normal function of the network, so this paper introduces deep learning to realize the mining of network security vulnerabilities. Deep learning learns and interprets the multimodal data features of distributed network security vulnerabilities by simulating the neurons of the human brain, so as to achieve the purpose of security vulnerability mining. The following mainly introduces the overall structure of the deep learning model and the process of distributed network vulnerability mining. The vulnerability mining model is mainly composed of input, hidden and output layers. The first is the input of the model, which accepts two inputs: the network vulnerability data features to be mined and the vulnerability feature templates. Among them, the vulnerability feature templates are mainly obtained from the identified vulnerability data, and each template represents a vulnerability type. Secondly, the hidden layer of the model is used for feature processing and vulnerability mining.

III. B. NSSA Technical Framework

In order to construct the basic framework of NSSA technology, Endsley model and Agent theory are introduced in the paper respectively.

III. B. 1) Endsley model

Generally speaking, Endsley model is applied in the data processing process of NSSA technology and plays a larger role in the construction of NSSA technology framework.

The workflow of the Endsley model is roughly as follows: understand the network security posture and obtain multiple factors in the network environment as inputs for situational awareness. Its precise description is shown in equation (16):

$$S(t)=\{C_1(t),\cdots,C_i(t)\cdots,C_n(t)\} \tag{16}$$

Where t denotes the current moment of the network state, n denotes the number of factors in the network environment, such that the symbols A,T and L denote the deterministic values, temporal attributes and spatial attributes that threaten the network security, and $C_i(t)$ denotes all the attributes of the i th influencing factor at the t th moment, which is mathematically expressed as:

$$C_i(t)=\{A,T,L\} \tag{17}$$

In the specific workflow, the Endsley model needs to accurately match the set of attributes with threats with the relevant historical data in order to extract the most influential situational elements [19]. Here let T denote the extraction function of situational awareness, X denotes the matching calculation method between the set of attributes with threats and the historical data, and X_{re} denotes the set of attributes with threats, which has the general form of $X_n=\{S(1),S(2),\cdots,S(k),\cdots\}$, and X_n denotes the network security historical data in the form of knowledge base and X_F denotes the final matching result in the form of feature vector, then this matching computational equation is shown in equation (18):

$$T:X_n\times X_H\rightarrow X_F \tag{18}$$

III. B. 2) Agent Theory

Generally speaking, Agent theory is a computer software system that can intelligently accomplish a specific goal in a specific application scenario.A-gent theory originates from the technical research of artificial intelligence, and is now widely used in a variety of fields such as distributed computing and software development.

III. B. 3) Radial Basis Neural Networks

Radial basis neural network consists of input layer, hidden layer and output layer, and its parameters are mainly the center, variance and weight of the basis function. In order to achieve data fusion with excellent functionality, in radial basis function network, this paper adopts the self-organized selection method as the main learning method in the organizational and supervised learning phase. Among them, the organizational learning stage is responsible for learning the center and variance parameters of the basis function, and the supervised learning stage is responsible for learning the weights parameters. Let the total number of clustering centers be I, the i th learning center is $t_i, i=1,\cdots,I$, and the center of the basis function is obtained in the n th iteration as $t_i(n)$, and its specific learning steps are as follows:

Step 1 initialization, so that the number of iterations $n=0$, randomly set I different initial learning centers $t_i(0)$.

Step 2 Randomly input the corresponding training data samples X, the

Step 3 search all I learning centers $t_i(j), j=0,\cdots,n$ such that the distance from training data sample X to the learning center is minimized.

Step 4 selects and substitutes new learning centers $t_i(n+1)$, using Eq. (19).

$$t_i(n+1)=\begin{cases}t_i(n)+\varphi[X(n)-t_i(n)], i=i(X)\\ t_i(n), i\neq i(X)\end{cases} \tag{19}$$

Step 5 detects whether all the training samples have been learned and whether the sample distribution is deterministic. If the samples are deterministic, the learning process ends. Otherwise, make $n = n + 1$ and turn to step 2.

Step 6 Output the final basis function center $t_i, i = 1, \cdots, I$.

III. B. 4) NSSA security assessment

Utilizing the training process of radial basis neural network, this paper obtains a neural network assessment method with strong assessment capability. The method, together with the real-time collected network state information, can realize the multi-source data assessment, early warning and decision-making of network security posture.

III. C. Multimodal Heterogeneous Network Data Security Assessment

III. C. 1) Test delay

In order to verify the effectiveness of the method in this paper, it is necessary to carry out simulation experimental tests, the specific experimental environment: Intel (R) Core (TM) i5 processor, main frequency of 2.0 GHz, 4 GB of memory. Experiments to select the private key can be recovered as well as pseudo-randomized bilinear mapping method as a comparative method, the division of several groups of experiments. (1) The experiments compare the test delay of the three test methods. The experimental comparison results are shown in Fig. 4, the delay of this paper's method is significantly lower, the average test delay in each test sample number is 2.555ms, which is better than the other two kinds of 6.873ms and 9.264ms, which is mainly due to the fact that this paper's method has been improved on the basis of the traditional method, which prompts the system to discover network security vulnerabilities in a timely manner and report to the police.

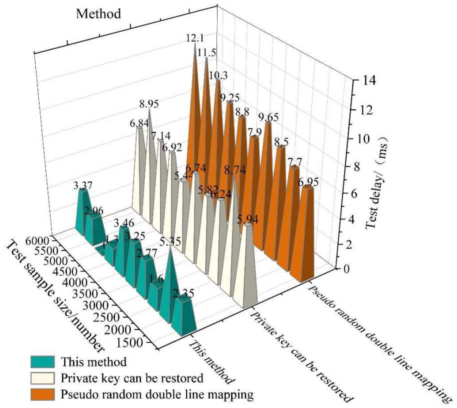
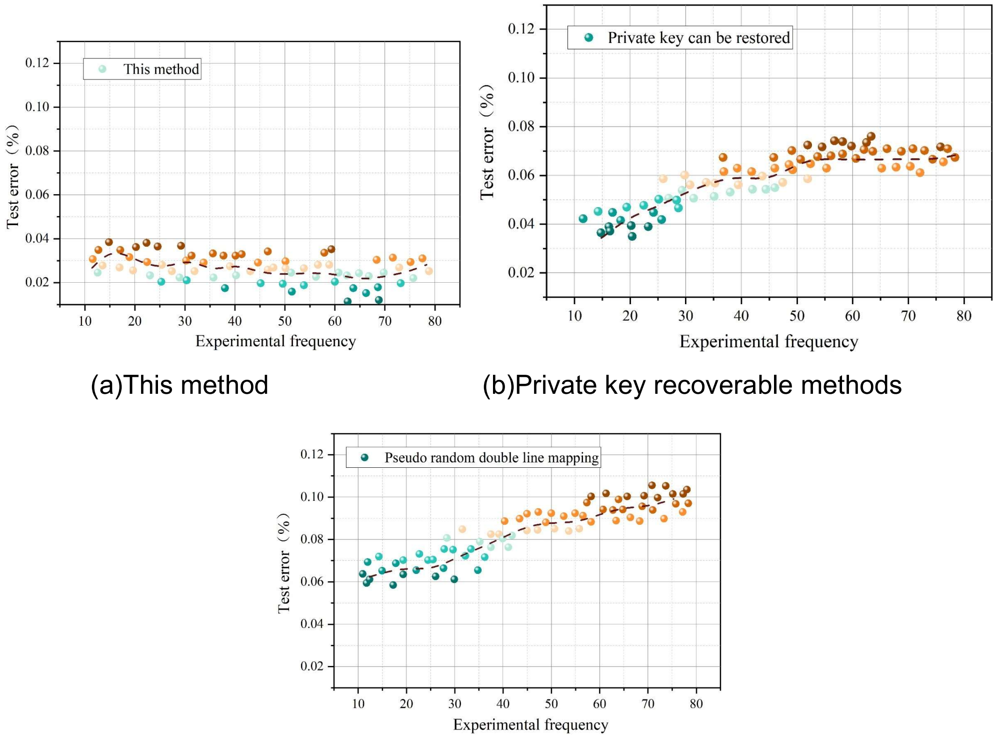


Figure 4: Test delay

III. C. 2) Testing errors

The experiment compares the accuracy of the three methods, Figure 5 shows the test error comparison results of different methods, Figure (a) is the method of this paper, Figure (b) is the private key recoverable method, and Figure (c) is the pseudo-random bilinear mapping method, the dotted line in the figure is the approximate trend of the error. The experiment takes the test error as an evaluation index, where the lower the test error, the better the test performance of the method, with the increase of the number of experiments, the test error of this paper's algorithm does not increase but decreases, and stabilizes in the interval of 0~0.04. The errors of the private key recoverable method and pseudo-random bilinear mapping method are between [0.03,0.08], [0.05,0.11], indicating that the method in this paper can obtain more satisfactory test results and effectively ensure the stable operation of multimodal information fusion for English vocabulary teaching.



(a)This method (b)Private key recoverable methods

(c)Pseudo-random two-wire mapping method

Figure 5: Test error comparison results of different methods

III. C. 3) Volume of data stored by read and write operations

In order to verify the effectiveness of the studied method for secure storage of multimodal heterogeneous network data in cloud computing environment, evaluate the effect of the studied method in practical application, and further verify its feasibility and effectiveness.

Multimodal heterogeneous network data security storage using the method of this paper, statistics on the amount of data stored per second when the method is used for read and write operations, the statistical results are shown in Fig. 6, the method of this paper for multimodal heterogeneous network data security storage in the artificial intelligence environment can ensure data security under the circumstances of large amounts of data storage, to meet the demand for reading and writing of large amounts of data in the artificial intelligence environment. The method in this paper can guarantee the data processing capacity of 2~2.1Mb/s, which has high storage efficiency.

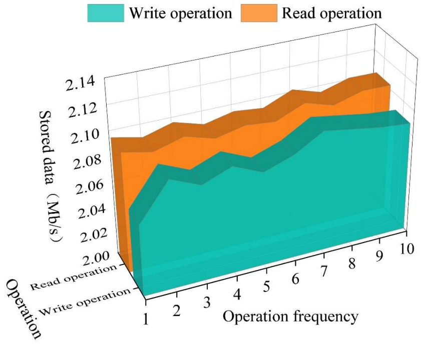


Figure 6: Read and write operations to store data

III. C. 4) Network throughput changes

The statistical results of the change in network throughput rate at different data volumes using the method of this paper for multimodal heterogeneous network data storage are shown in Fig. 7. Using the method of this paper for multimodal heterogeneous network data security storage, different data volumes are able to maintain a throughput higher than 3Mb/s, with an average of 5.506Mb/s. The method in this paper utilizes radial basis neural network for data processing, which effectively improves the data storage performance.

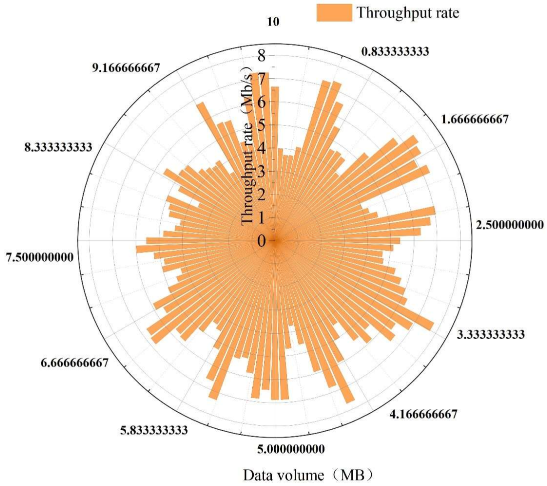
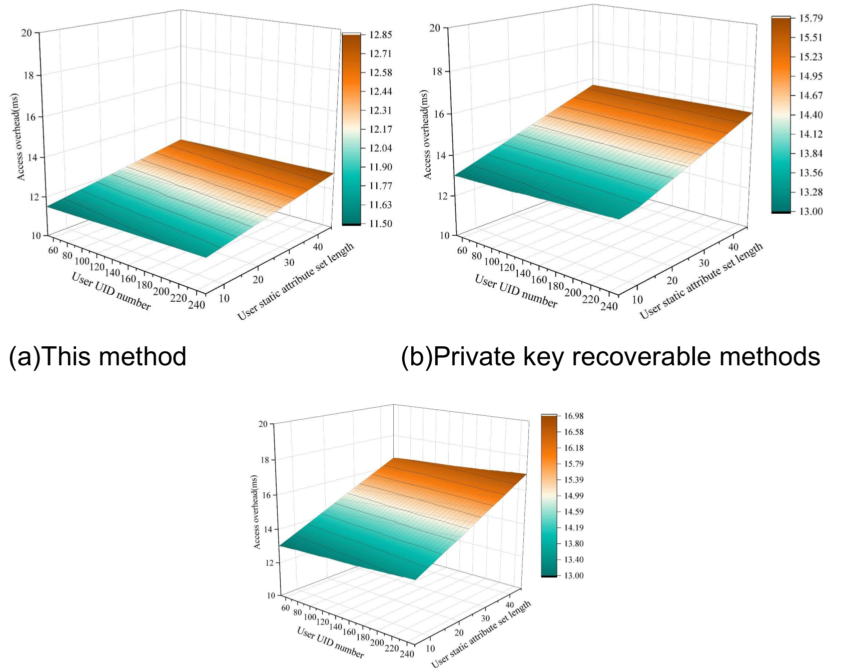


Figure 7 Throughput change of multi-mode heterogeneous network

III. C. 5) Access Overhead Detection

Fig. 8 shows the access overhead detection results, Fig. (a) shows the method of this paper, Fig. (b) shows the private key recoverable method, and Fig. (c) shows the pseudo-random bilinear mapping method, with the increase of the number of user UIDs and the length of the user's static attribute set, the access overhead of the proposed method, the private key recoverable method and the pseudo-random bilinear mapping method increase, but the increasing trend of the proposed method is more gentle, and the access overhead increases from 11.5ms to 12.85ms, which is lower than that of the private key recoverable method and pseudo-random bilinear mapping method and always stays below 13ms.



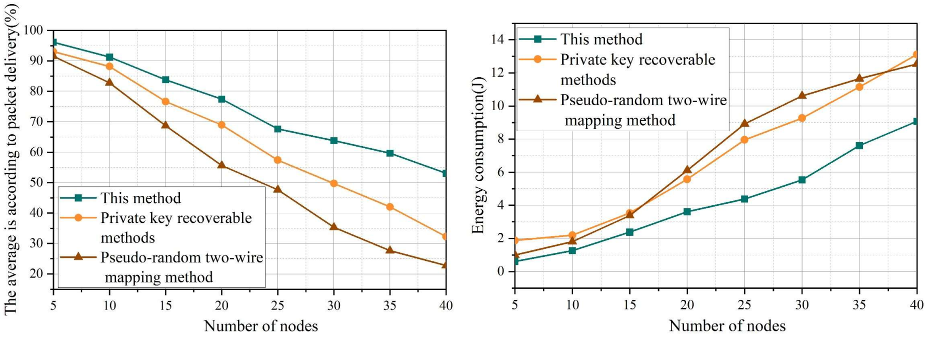
(a)This method (b)Private key recoverable methods

(c)Pseudo-random two-wire mapping method

Figure 8: Access overhead test results

III. C. 6) Average energy consumption test results

In order to further test the performance of the three methods in cross-layer secure access to data in multimodal heterogeneous networks, any node in the multimodal heterogeneous network is selected as the target node, so that the node moves at a constant speed of 5 m/s, and detects the average packet delivery rate and the average energy consumption of the proposed method, the private-key recoverable method, and the pseudo-random bilinear mapping method under the increase of the target node, and the higher the average packet delivery rate and the lower the average energy consumption, the stronger the cross-layer access performance of the corresponding methods. Energy consumption, the stronger the cross-layer access performance of the corresponding method. The experimental results are shown in Fig. 9, Fig. (a) shows the average packet delivery rate detection results, and Fig. (b) shows the average energy consumption detection results. In the process of increasing the number of nodes, the average packet delivery rate of the three methods decreases and the average energy consumption increases, but the average packet delivery rate of the proposed method is higher than that of the other two methods throughout the process, and the average energy consumption is always lower than that of the other two methods, with the average values of the packet delivery rate and the average energy consumption of 74.102% and 4.308J, respectively, and the gap is more and more obvious with the increase of the number of nodes, indicating that the proposed method has a better performance in multi-layer access than the other two methods, and the average energy consumption of the proposed method is lower than the other two methods. The average values of packet delivery rate and average energy consumption are 74.102% and 4.308J, respectively, and the gap increases with the increase of the number of nodes, which indicates that the proposed method has better performance in cross-layer secure access of multimodal heterogeneous network data.



(a)The average is reported by packet delivery rate

(b)Average energy test results

Figure 9: Safe access performance

IV. Conclusion

In this paper, we construct an English vocabulary teaching model based on multimodal information fusion from the four levels of target encoder, decoder, knowledge point embedding, and modal fusion in turn, and design research experiments to evaluate the effect of English vocabulary teaching with multimodal information fusion and the security of heterogeneous network data respectively.

Comparing the English vocabulary learning attitudes of the experimental class and the control class, the mean value of motivation in class A is 3.214, and the mean value of motivation in class B is 2.469. At the same time, through the independent samples t-test, it can be seen that under the assumption of equal variance, the t-value of motivation in the two classes is 5.166, and the Sig value is 0.000, which is less than 0.05, and class A is significantly higher than class B. This indicates that the multimodal information integration teaching model is conducive to improving students' motivation and positively affecting vocabulary learning.

Comparing this paper's algorithm with the private key recoverable method and the pseudo-random bilinear mapping method, the average test delay of this paper's method is 2.555ms for each number of test samples, which is better than the other two's 6.873ms and 9.264ms, and this paper's method's delay is obviously lower, and the test error decreases with the increase of the number of experiments, and stabilizes in the interval of 0~0.04, which ensures the stable operation of the English vocabulary stable operation of multimodal information fusion for teaching.

References

[1] Kim, J., Lee, H., & Cho, Y. H. (2022). Learning design to support student-AI collaboration: Perspectives of leading teachers for AI in education. Education and information technologies, 27(5), 6069-6104.

[2] Ruiz-Rojas, L. I., Acosta-Vargas, P., De-Moreta-Llovet, J., & Gonzalez-Rodriguez, M. (2023). Empowering education with generative artificial intelligence tools: Approach with an instructional design matrix. Sustainability, 15(15), 11524.

[3] Zhao, Y., & Yang, Z. (2024). Research on collaborative innovation optimization strategies for digitally enabled higher education ecosystems. Plos one, 19(4), e0302285.

[4] Choi, H. (2024). Effects of Digital Technology-Integrated English Classes on Middle School Students' Vocabulary. Multimedia-Assisted Language Learning, 27(4).

[5] Hay, L. (2019). In the Classroom with Multi-Modal Teaching. Teaching Classics with Technology. London: Bloomsbury, 229-238.

[6] Wang, M. (2022). The cultivation of students’ multiple practical ability under the multimodal blended teaching mode. International Journal of Emerging Technologies in Learning (iJET), 17(12), 106-120.

[7] Schüler, A. (2019). The integration of information in a digital, multi-modal learning environment. Learning and instruction, 59, 76-87.

[8] Durongbhandhu, N., & Suwanasilp, D. (2021). Effectiveness of Multimodal Glossing Reading Program on English Vocabulary Acquisition. English Language Teaching, 14(6), 62-75.

[9] Luo, T., Cai, N., Li, Z., Miao, J., Pan, Z., Shen, Y., ... & Zhang, M. (2021, March). MagicChem: A Multi-modal Mixed Reality System Based on Needs Theory for Chemical Education. In 2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW) (pp. 544-545). IEEE.

[10] Choi, J., & Yi, Y. (2016). Teachers' integration of multimodality into classroom practices for English language learners. Tesol Journal, 7(2), 304-327.

[11] Sutrisno, D., Abidin, N. A. Z., Pambudi, N., Aydawati, E., & Sallu, S. (2023). Exploring the benefits of multimodal literacy in English teaching: Engaging students through visual, auditory, and digital modes. Global Synthesis in Education Journal, 1(2), 1-14.

[12] Suwastini, N. K. A., Marantika, P. D., Adnyani, N. L. P. S., Mandala, M. A. K., & Artini, N. N. (2021). Multimodal teaching in EFL context: A literature review. Edu-Ling: Journal of English Education and Linguistics, 4(2), 140-151.

[13] Kummin, S., Surat, S., Kutty, F. M., Othman, Z., & Muslim, N. (2020). The use of multimodal texts in teaching English language oral skills. Universal Journal of Educational Research, 8(12), 7015-7021.

[14] Hafner, C. A. (2020). Digital Multimodal Composing: How to Address Multimodal Communication Forms in ELT. English teaching, 75(3), 133-146.

[15] Mu, H. (2018). A study on English acquisition from the perspective of the multimodal theory. Theory and Practice in language Studies, 8(6), 618-623.

[16] Cárcamo, M. M. A., Cartes, R. A. C., Velásquez, N. E. E., & Larenas, C. H. D. (2016). The impact of multimodal instruction on the acquisition of vocabulary. Trabalhos em Linguística Aplicada, 55(1), 129-154.

[17] Guo Baicang,Liu Hao,Yang Xiao,Cao Yuan,Jin Lisheng & Wang Yinlin. (2025). Multi-modal information fusion for multi-task end-to-end behavior prediction in autonomous driving. Neurocomputing,634,129857-129857.

[18] Shenrong Lv,Bo Yang,Ruiyang Wang,Siyu Lu,Jiawei Tian,Wenfeng Zheng... & Lirong Yin. (2024). Dynamic Multi-Granularity Translation System: DAG-Structured Multi-Granularity Representation and Self-Attention. Systems,12(10),420-420.

[19] Bhatia Manini R.,Malhotra Atul,Bansal Utkarsh,Singh Jai Vir & Kumar Arunaz. (2022). Using the Endsley Model to Evaluate Simulation-Based Situation Awareness Training to Medical and Nursing Students in India: A Qualitative Analysis. Simulation in Healthcare: The Journal of the Society for Simulation in Healthcare,